# Music Genre Classification

Mark Gameng

# Motivation + Problem Description


- Spotify have listed over 50 million songs and over 40 thousand new songs are added every day
- Increasing interests towards Music Information Retrieval (MIR)
- Lots of work done in Music Genre Classification

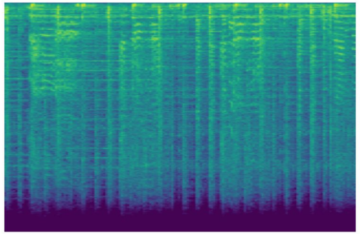
## Genre Classification:

- Use extracted features from audio to accurately classify the genre

Table 1: Results on GTZAN

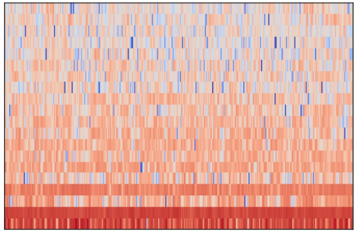
| Model | Feature | Accuracy |
|---|---|---|
| MusicRecNet + SVM [1] | Melspectrogram | 97.6% |
| Broadcast NN [2] | Melspectrogram | 93.9% |
| CNN + 1-layer RNN [8] | STFT | 90.2% |
| CNN + 2-layer RNN [8] | STFT | 88.8% |
| CNN [8] | STFT | 88.0% |
| 2-layer BRNN [4] | STFT | 76.2% |

# Representations of Music



(a) STFT

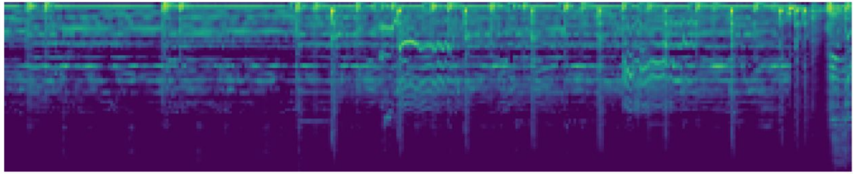
- Short-time Fourier Transform
- Sinusoidal frequency and phase content of local sections of a signal as it changes over time
- Process: Divide longer time segment into shorter equal segments and compute the Fourier transform separately on each segment. This shows the Fourier spectrum on each segment and then together plotted as a function of time



(b) MFCC

- Mel-frequency cepstral coefficients
- MFCCs are coefficients that make up a Mel-frequency cepstrum (MFC)
- Derived from a type of cepstral representation of audio
- Approximates the human auditory system response more closely than others and allows for a better representation of sound.

# Representations of Music cont.



(c) Melspectrogram

- Essentially a spectrogram (frequency vs time) in mel scale
- Mel scale is a logarithmic transformation of a signal's frequency
- Mel scale basically mimics our own perception of sound

# Data

**GTZAN [7]**

- 10 genres; 100 songs of 30 seconds length for each; Total: 1000
- Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae, Rock

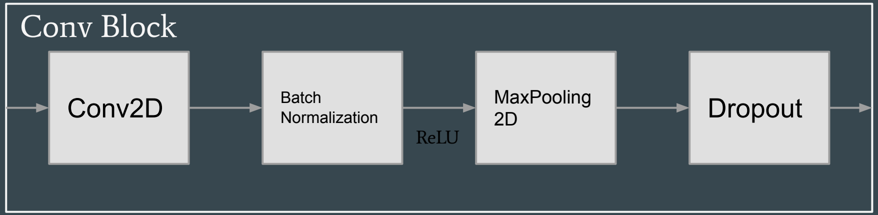
**Extended Ballroom [3]**

- 13 genres (4 removed for < 70 samples); 250 - 530 songs for each; Total: 3,992
- Chacha, Jive, Quickstep, Rumba, Samba, Tango, Viennesewaltz, Waltz, Foxtrot

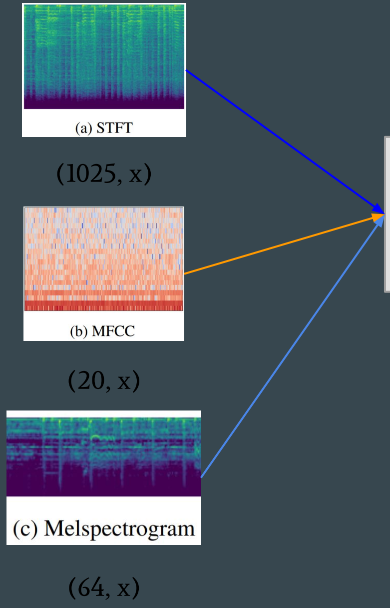
5, 8, 10 segments
6 sec, 3.75 sec, 3 sec
5,000: 8,000: 10,000 samples | 19,960: 31,936: 39,920 samples

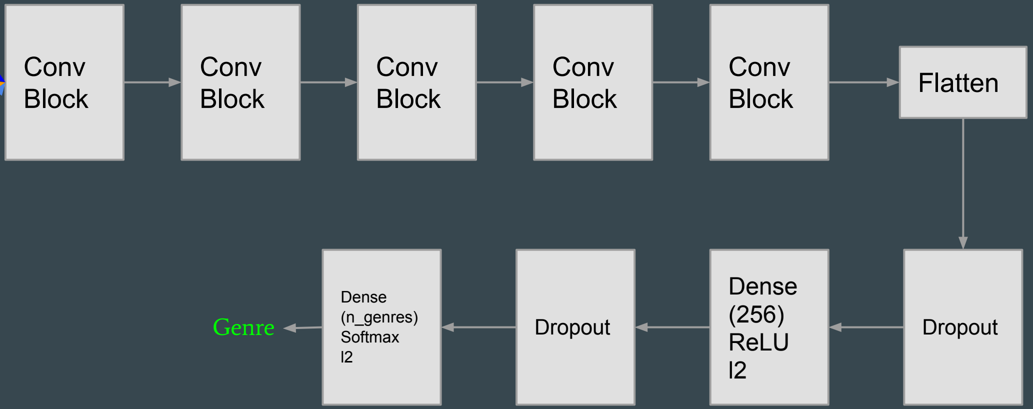
64% Train | 16 % Validation | 20 % Test

# Methodology - CNN




Due to different inputs (sizes), models have different but similar layouts



# CNN Results

## GTZAN

| Feature | Split | Test Accuracy |
|---|---|---|
| STFT | 8 | 0.863125 |
| STFT | 5 | 0.855000 |
| MELSPECTROGRAM | 8 | 0.840000 |
| MELSPECTROGRAM | 10 | 0.833000 |
| STFT | 10 | 0.825000 |
| MELSPECTROGRAM | 5 | 0.815000 |
| MFCC | 10 | 0.759000 |
| MFCC | 8 | 0.701250 |
| MFCC | 5 | 0.610000 |

## Extended Ballroom

| Feature | Split | Test Accuracy |
|---|---|---|
| STFT | 5 | 0.903808 |
| STFT | 8 | 0.903413 |
| STFT | 10 | 0.894664 |
| MELSPECTROGRAM | 5 | 0.894289 |
| MELSPECTROGRAM | 8 | 0.870539 |
| MELSPECTROGRAM | 10 | 0.850576 |
| MFCC | 5 | 0.845190 |
| MFCC | 8 | 0.815435 |
| MFCC | 10 | 0.805361 |

## Comparisons

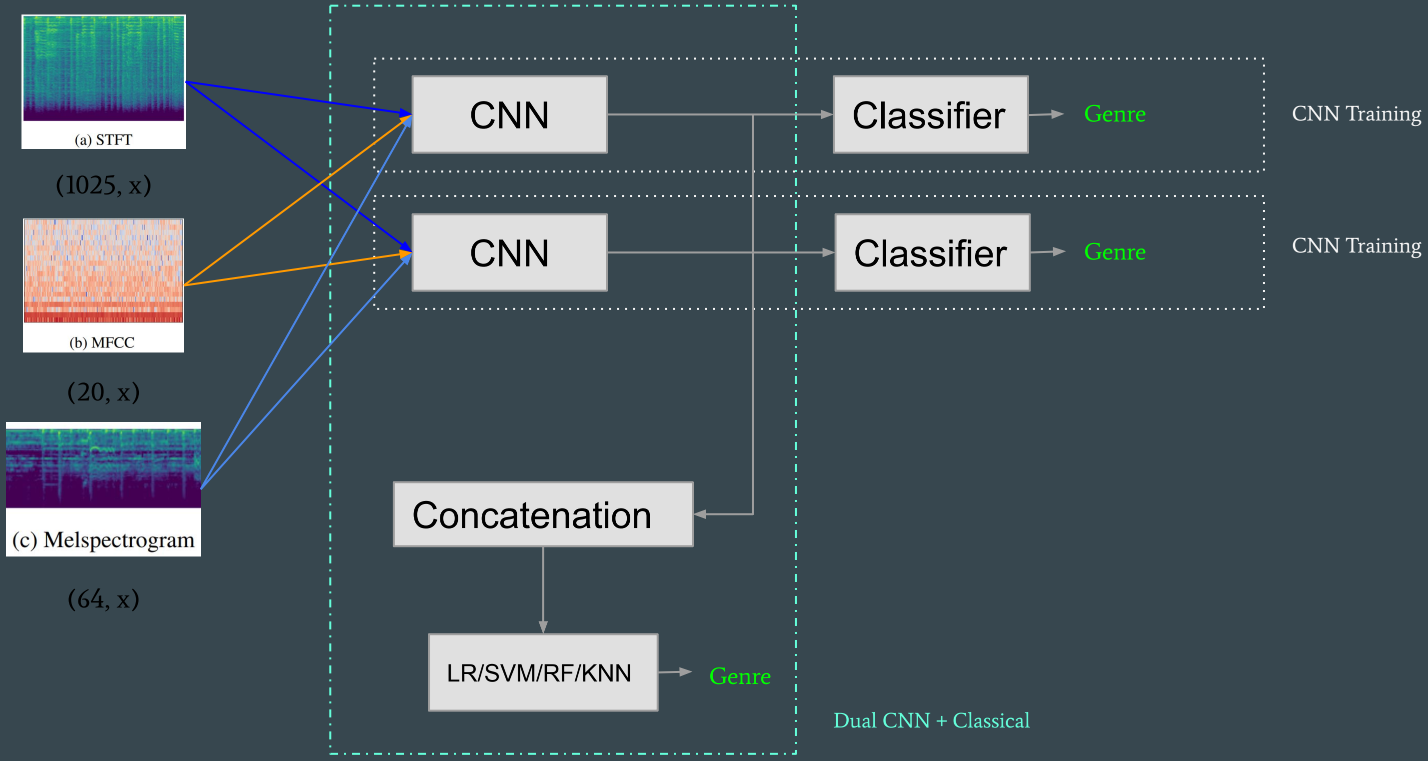
Table 1: Results on GTZAN

| Model | Feature | Accuracy |
|---|---|---|
| MusicRecNet + SVM [1] | Melspectrogram | 97.6% |
| Broadcast NN [2] | Melspectrogram | 93.9% |
| CNN + 1-layer RNN [8] | STFT | 90.2% |
| CNN + 2-layer RNN [8] | STFT | 88.8% |
| CNN [8] | STFT | 88.0% |
| **CNN - 8 split** | **STFT** | **86.3%** |
| **CNN - 8 split** | **Melspectrogram** | **84.0%** |
| 2-layer BRNN [4] | STFT | 76.2% |
| **CNN - 8 split** | **MFCC** | **70.1%** |

Table 2: Results on Extended Ballroom

| Model | Feature | Accuracy |
|---|---|---|
| Broadcast NN [2] | Melspectrogram | 97.2% |
| CNN + 1-layer RNN [8] | STFT | 92.3% |
| CNN + 2-layer RNN [8] | STFT | 92.5% |
| CNN [8] | STFT | 92.2% |
| 2-layer BRNN [4] | STFT | 90.3% |
| **CNN - 8 split** | **STFT** | **90.3%** |
| **CNN - 8 split** | **Melspectrogram** | **87.0%** |
| **CNN - 8 split** | **MFCC** | **81.5%** |

# Methodology - Dual CNN + Classical



# Dual CNN + Classical Results

## GTZAN

| Method | Feature | Split | Test Accuracy |
|---|---|---|---|
| CNN + KNN | STFT, MELSPECTROGRAM | 8 | 0.958125 |
| CNN + KNN | STFT, MFCC | 8 | 0.953125 |
| CNN + SVM | STFT, MELSPECTROGRAM | 8 | 0.950625 |
| CNN + SVM | STFT, MFCC | 8 | 0.940000 |
| CNN + LR | STFT, MELSPECTROGRAM | 8 | 0.932500 |
| CNN + LR | STFT, MFCC | 8 | 0.921250 |
| CNN + RF | STFT, MFCC | 8 | 0.916875 |
| CNN + RF | STFT, MELSPECTROGRAM | 8 | 0.916250 |
| CNN | STFT, MFCC | 8 | 0.908750 |
| CNN | STFT, MELSPECTROGRAM | 8 | 0.905625 |
| CNN + KNN | MFCC, MELSPECTROGRAM | 8 | 0.897500 |
| CNN | MFCC, MELSPECTROGRAM | 8 | 0.891875 |
| CNN + SVM | MFCC, MELSPECTROGRAM | 8 | 0.891250 |
| CNN + LR | MFCC, MELSPECTROGRAM | 8 | 0.869375 |
| CNN + RF | MFCC, MELSPECTROGRAM | 8 | 0.853125 |

## Extended Ballroom

| Method | Feature | Split | Test Accuracy |
|---|---|---|---|
| CNN + SVM | STFT, MFCC | 8 | 0.924076 |
| CNN + RF | STFT, MFCC | 8 | 0.919693 |
| CNN + SVM | STFT, MELSPECTROGRAM | 8 | 0.919537 |
| CNN + KNN | STFT, MFCC | 8 | 0.918284 |
| CNN + RF | STFT, MELSPECTROGRAM | 8 | 0.915780 |
| CNN + LR | STFT, MFCC | 8 | 0.914684 |
| CNN + KNN | STFT, MELSPECTROGRAM | 8 | 0.914058 |
| CNN | STFT, MFCC | 8 | 0.913118 |
| CNN + LR | STFT, MELSPECTROGRAM | 8 | 0.912805 |
| CNN | STFT, MELSPECTROGRAM | 8 | 0.911240 |
| CNN + SVM | MFCC, MELSPECTROGRAM | 8 | 0.898403 |
| CNN | MFCC, MELSPECTROGRAM | 8 | 0.891202 |
| CNN + LR | MFCC, MELSPECTROGRAM | 8 | 0.889167 |
| CNN + RF | MFCC, MELSPECTROGRAM | 8 | 0.883688 |
| CNN + KNN | MFCC, MELSPECTROGRAM | 8 | 0.877113 |

## Comparisons

Table 1: Results on GTZAN

| Model | Feature | Accuracy |
|---|---|---|
| MusicRecNet + SVM [1] | Melspectrogram | 97.6% |
| **Dual CNN + KNN** | **STFT, Melspectrogram** | **95.8%** |
| **Dual CNN + KNN** | **STFT, MFCC** | **95.3%** |
| **Dual CNN + SVM** | **STFT, MFCC** | **94.0%** |
| Broadcast NN [2] | Melspectrogram | 93.9% |
| CNN + 1-layer RNN [8] | STFT | 90.2% |
| CNN + 2-layer RNN [8] | STFT | 88.8% |
| CNN [8] | STFT | 88.0% |
| **CNN** | **STFT** | **86.3%** |
| **CNN** | **Melspectrogram** | **84.0%** |
| 2-layer BRNN [4] | STFT | 76.2% |
| **CNN** | **MFCC** | **70.1%** |

Table 2: Results on Extended Ballroom

| Model | Feature | Accuracy |
|---|---|---|
| Broadcast NN [2] | Melspectrogram | 97.2% |
| CNN + 2-layer RNN [8] | STFT | 92.5% |
| **Dual CNN + SVM** | **STFT, MFCC** | **92.4%** |
| CNN + 1-layer RNN [8] | STFT | 92.3% |
| CNN [8] | STFT | 92.2% |
| **Dual CNN + RF** | **STFT, MFCC** | **91.9%** |
| **Dual CNN + SVM** | **STFT, Melspectrogram** | **91.9%** |
| 2-layer BRNN [4] | STFT | 90.3% |
| **CNN** | **STFT** | **90.3%** |
| **CNN** | **Melspectrogram** | **87.0%** |
| **CNN** | **MFCC** | **81.5%** |

# Sources

[1] Elbir, A., & Aydin, N. (2020). Music genre classification and music recommendation by using deep learning. *Electronics Letters*, *56*(12), 627-629.

[2] Liu, C., Feng, L., Liu, G., Wang, H., & Liu, S. (2021). Bottom-up broadcast neural network for music genre classification. *Multimedia Tools and Applications*, *80*(5), 7313-7331.

[3] Marchand, U., & Peeters, G. (2016). The extended ballroom dataset.

[4] Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673-2681.

[7] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, *10*(5), 293-302.

[8] Yang, R., Feng, L., Wang, H., Yao, J., & Luo, S. (2020). Parallel recurrent convolutional neural networks-based music genre classification method for mobile devices. *IEEE Access*, *8*, 19629-19637.

# Demo


```
['rock', 'disco', 'disco', 'disco', 'rock', 'disco', 'disco', 'rock', 'disco', 'hiphop', 'pop', 'rock', 'pop', 'pop', 'pop
', 'pop', 'disco', 'pop', 'pop', 'disco', 'disco', 'rock', 'rock', 'disco', 'hiphop', 'rock', 'country', 'pop', 'pop', 'pop
', 'disco', 'rock', 'pop', 'disco', 'reggae', 'hiphop', 'rock', 'hiphop', 'disco', 'disco', 'hiphop', 'disco', 'hiphop', 'd
isco', 'disco', 'disco', 'disco', 'disco', 'pop', 'disco', 'rock', 'hiphop', 'pop', 'hiphop', 'disco', 'disco', 'disco', 'h
iphop', 'hiphop', 'hiphop', 'country', 'disco', 'disco', 'rock']

Summary:
{'rock': 11, 'disco': 26, 'hiphop': 11, 'pop': 13, 'country': 2, 'reggae': 1}

Predicted: disco
```

# Demo cont.

```
['rumba', 'waltz', 'waltz', 'waltz', 'waltz', 'waltz', 'waltz', 'rumba', 'rumba', 'viennesewaltz', 'quickstep', 'tango', 't
ango', 'tango', 'tango', 'tango', 'tango', 'tango', 'waltz', 'tango', 'waltz', 'samba', 'foxtrot', 'quickstep', 'tango', 't
ango', 'waltz', 'quickstep', 'tango', 'tango', 'tango', 'tango', 'tango', 'jive', 'foxtrot', 'tango', 'tango', 'viennesewal
tz', 'jive', 'tango', 'viennesewaltz', 'tango', 'tango', 'viennesewaltz', 'tango', 'tango', 'tango', 'tango']

Summary:
{'rumba': 3, 'waltz': 9, 'viennesewaltz': 4, 'quickstep': 3, 'tango': 24, 'samba': 1, 'foxtrot': 2, 'jive': 2}

Predicted: tango
```


The End